ARI Research Note 99-11

Data Base Documentation Standards for Extant Datasets

Ani S. DeFazio and Winnie Y. Young
Human Resources Research Organization

Organization and Personnel Resources Research Unit
Paul A. Gade, Chief

January 1999

U.S. Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the U.S. Total Army Personnel Command

EDGAR M. JOHNSON
Director

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Ronald B. Tiggle

NOTICES

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 1999	Final August 1996 - December 1996

4. TITLE AND SUBTITLE
data base documentation standards for extant datasets

5. FUNDING NUMBERS
MDA903-93-D-0032 DO 0051
6900C6
62785A790
63007A792

6. AUTHOR(S)
Ani S. DiFazio and Winnie Y. Young

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 400
Alexandria, Virginia 22314

8. PERFORMING ORGANIZATION REPORT NUMBER
FR-EADD-97-05

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: TAPC-ARI-RP
5001 Eisenhower Avenue
Alexandria, Virginia 22333-5600

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARI Research Note 99-11

11. SUPPLEMENTARY NOTES
Contracting Officer's Representative: Ronald B. Tiggle

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Since 1975, the U.S. Army Research Institute (ARI) has collected a wide array of manpower Personnel Research (MPR) data in support of its research activities. Until this current effort, there have been no formal procedures or guidelines for the documentation and archive of these numerous databases. The ability of new users to access and utilize extant ARI data, whether collected by ARI staff or by outside contractors, is heavily dependent on the knowledge of those ARI staff members who worked most closely with the data. With organizational turnover and downsizing, critical information needed to access and use data by new users will be lost over time. As Phase I of a two phase effort, the Human Resources Research Organization (HumRRO), and Fu Associates were awarded a contract to develop standards for documentation and archive of extant ARI datasets. The development of these documentation and archive standards is the subject of this report.

14. SUBJECT TERMS
dataset
data base
documentation
data file
archive

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Acknowledgements

The authors acknowledge and appreciate the helpful cooperation from various individuals involved in this project. We would like to thank Dr. Ronald Tiggle, who has served as Technical Monitor of this Delivery Order. We would also like to thank Dianne Driessen, Fumiyo Tao, and Donna Peck of Fu Associates who provided critical information and insight in the development of the baseline documentation instrument.

TABLE OF CONTENTS

Introduction 1

Objective 1

Review of Existing Data Documentation Standards and Guidelines 1

Technical Approach 4
- Assessing Dataset Characteristics 4
- Documentation Requirement Levels and Standards 5
- Confidentiality 12
- Archive Media Evaluation 12

Conclusions 15

APPENDIX A A-1

APPENDIX B B-1

APPENDIX C C-1

APPENDIX D D-1

Data Base Documentation Standards for Extant Datasets

Introduction

Since 1975, the U.S. Army Research Institute (ARI) has collected a wide array of Manpower Personnel Research (MPR) data in support of its research activities. Until this current effort, there have been no formal procedures or guidelines for the documentation and archiving of these numerous databases. As a result, many extant datasets are neither currently documented nor archived at a central location. The ability of new users to access and utilize extant ARI data, whether collected by ARI staff or by outside contractors, is heavily dependent on the knowledge of those ARI staff members who worked most closely with the data. With organizational turnover and downsizing, critical information needed to access and use data by new users will be lost over time.

An effort was made in 1992 to develop basic documentation of ARI MPR datasets (Holway, Tao, Ramsey, Payne, & Haupt, 1992). It is now four years later and there are new datasets to be cataloged and updates may be required for some previously cataloged datasets. Also, the Haloway et al. document entitled *Catalog and Assessment of the Manpower and Personnel Research Division Data Bases* was an initial effort. ARI recognizes the need for a more extensive effort to document and archive both extant and future ARI datasets.

Objective

The objective of this document is to establish documentation and archive standards for extant MPR datasets. We begin by conducting a literature review of industry standards for data documentation, primarily in the area of social science research. It is important to note that we have focused our review on the most relevant category of data and research, since vastly different types (for example, bibliographic, geological, logistical) are likely to require varying documentation and archiving approaches.

Review of Existing Data Documentation Standards and Guidelines

In contemporary society, almost every industry and area of endeavor is involved in the development, use, and maintenance of numerous and often complex automated data files and information systems. Among those involved in social science inquiry and research, there is no question of the usefulness and value in the proper documentation and archive of data.

There are a number of benefits associated with careful data documentation. In its *Guide to Social Science Data Preparation and Archiving,* the Inter-university Consortium for Political and Social Research (ICPSR)reviews Fienberg's (1994) cost-benefit approach. Specifically, this approach argues that data documentation and archiving:

- Reinforces open scientific inquiry
- Encourages diversity of analysis and opinions
- Promotes new research and allows for the testing of new or alternative methods
- Improves methods of data collection and measurement through the scrutiny of others
- Reduces costs by avoiding duplicate data collection efforts
- Provides an important resource for training in research

Many of these benefits are echoed by the American Psychological Association (APA, 1992), the American Statistical Association (ASA, 1993), the Department of Defense (DoD, 1994, 1996; DMDC, 1996), and other governmental and quasi-governmental agencies (U.S. Department of Justice, 1996; Roistacher, 1980; Peterson & Corella, 1991).

Regarding the cost component of the cost-benefit model of data documentation, Fienberg suggests that costs of archiving and documentation must be balanced against costs that can, on occasion, arise when data are not documented or are documented poorly. Some examples of such costs include:

- Responding to the discovery of errors or assertions of errors by others
- Dealing with criticism of the study based on other investigators' analyses
- Foregoing rights to discoveries that others may make later
- Dealing with breaches of confidentiality by other investigators.

In the final analysis, the industry consensus is that the benefits of careful data archiving far outweigh the costs, both financial and otherwise.

Apart from universal acknowledgement of the overall value of documentation, what is clear from the review of the literature is that there is no one documentation standard that is applicable to all data. This is particularly true in the case of extant data. There is a vast array of technical characteristics of extant data that not only typify a particular dataset, but which, in turn, suggest documentation and archiving strategies.

Technical characteristics are objective physical attributes of datasets and include such factors as file type (e.g., EBCDIC, ASCII, or SAS system file), file structure (e.g., hierarchical or rectangular), the size of the dataset, and storage media. Many technical dataset characteristics should be decided before the dataset is actually constructed (ICPSR, 1996) and set in place during dataset creation and analysis. For example, issues relating to file structure, dataset and data element naming conventions, and data integrity are often best set forth in a data management plan, such as that produced for the ARI Project A/Career Force research program (Wise & Wang, 1983).

Unfortunately, the time for such careful planning and execution is long past for extant datasets. It is inherently the nature of extant data that these characteristics either cannot now be changed (e.g., the collection instruments used) or can be changed only at a substantial cost or effort (e.g., transformation of a complex SAS dataset to an ASCII file). It is clear that because the current task is to establish documentation standards for extant datasets, any schema that promulgates such standards must be responsive to the particular technical characteristics of the dataset in question.

In addition to technical data characteristics, the standard of documentation an extant dataset is subjected to will depend on pragmatic considerations and an assessment of the cost and benefit of thorough documentation. An example of pragmatic considerations would include the physical availability of the dataset to be documented. Pragmatic considerations also play an obvious part in cost/benefit assessments. For example, it could be argued that subjecting a ten year old dataset that will never be used again to the highest standards of documentation may be an unnecessary expenditure of government resources. This issue is discussed further later in this report.

In spite of a wide array of dataset characteristics that can influence documentation strategies, industry standards seem to suggest the following:

- data should be preserved as ASCII/EBCDIC files or portable transport system files such as SAS or SPSS files
- confidentiality of units of analysis must often be preserved
- any variables that link two or more files should be clearly identified
- the meaning of all variable codes should be clearly delineated
- codes used for missing values should be clearly defined

- constructed variables (including sampling weights, if any) should be clearly identified and defined
- methods for generating sampling weights, if any, and their appropriate use should be addressed
- data collection instrument(s) and sampling designs, where appropriate, should be included as part of dataset documentation
- recording media and formats should be identified (e.g., magnetic tape standards).

Technical Approach

Because documentation standards for extant datasets are inexorably intertwined with technical data characteristics and assessments of relative benefits and costs associated with documentation, we have developed a classification of requirement levels and associated documentation standards within those levels. The documentation requirement levels and associated standards are described in detail below. This system is predicated on the principal that the value of documentation increases with the frequency with which the dataset will be used in the future (or the likelihood that it will be used at all), as well as the overall importance or significance of the dataset in future efforts. Consequently, the greater the value of documentation, the more an investment in documentation is justified. However, it is important to remember that a judgment of the importance of thorough documentation may be distinct from its feasibility.

Assessing Dataset Characteristics

Our objective is to generate an appropriate level of documentation for each extant MPR dataset. To accomplish this goal, we first developed a comprehensive list of datasets[1]. In addition to providing a useful inventory of MPR datasets, the original plan was to use this list to select a subset of datasets on which to focus our attention on this effort. However, the scope of the effort widened as work progressed and now encompasses all identified MPR datasets.

[1] A dataset is defined here as a single physical file. Its use here conforms with the industry's typical use of the term, which is often used interchangeably with the term "data file." It has been our experience, however, that ARI research staff use these terms somewhat differently. In addition to a single physical file, ARI researchers use the term "dataset" to also refer to a collection of conceptually grouped albeit physically distinct files. For example, a collection of longitudinal files where only data collection year differentiates the physical files is typically called a "dataset" by ARI staff. To minimize confusion, we used these terms in our interactions with ARI researchers as they do. For example, our use of the terms "dataset" and "data file" conform to ARI useage in the ARI Research Dataset Questionnaire in Appendix C.

We started with an initial list of MPR datasets based on datasets recorded in the Holway et al. (1992) *Catalog* and the Sample of ARI Databases (ARI, 1995). This initial list of datasets was the topic of a brief survey of ARI staff designed to solicit additional names of datasets and information on dataset utility and importance[2]. The survey was also designed to identify the dataset Point of Contact (POC) - the ARI staff person most familiar with the dataset. A copy of this survey is presented in Appendix A. The additional datasets listed by respondents to this initial survey were added to the initial list of datasets to produce a more comprehensive list of MPR datasets. The addition, and in some cases, deletion of datasets from the list continued until approximately one month before the end of Phase I. In part, this was due to fact that when asked to provide dataset names, ARI researchers often provided the project or data base name. Some ARI datasets are part of a larger family of files or data base. For example, the Project A data base comprises numerous individual datasets. Because data bases can comprise varying numbers of datasets, this effort has focused on datasets, not data bases. Therefore, a considerable amount of Phase I time and effort was spent in identifying datasets among the information provided by ARI researchers.

Once MPR datasets were identified, we then began the task of classifying them into documentation requirement levels. Dataset POCs were asked to recommend a documentation requirement level for each dataset to which they were assigned by responding to a brief questionnaire. This questionnaire is presented in Appendix B.

Documentation Requirement Levels and Standards

Given industry standards and our current knowledge of extant MPR datasets, we have formed a three level system of documentation based on requirements of extant datasets. This classification system is designed to be hierarchical, so that all datasets will be documented at the highest level assigned, as well as all lower levels. This means that all datasets will be documented at the lowest level (i.e., Level I documentation) as well as at any assigned higher level(s). The following table summarizes the documentation requirements at each of the three levels.

Table 1. Documentation Requirement Levels

Requirements	Level I	Level II	Level III
Conduct Baseline Documentation	Yes	Yes	Yes
Collect Research Materials/Publications	No	Yes	Yes
Develop Comprehensive Codebook	No	No	Yes

[2] We collected information on dataset importance and utility since our original intention was to select a subset of datasets from the comprehensive list based on those characteristics. The focus of the project was later widened to encompass all identified MPR datasets.

The following section describes each of the three documentation requirement levels in detail.

Level I documentation. Level I documentation will be the minimum level of documentation for all extant MPR datasets. At this level, we are collecting basic technical dataset information. This information is collected through an in-depth survey called the ARI Research Dataset Questionnaire and is recorded electronically as an ACCESS file; it is presented in Appendix C. The current questionnaire is a modification and reorganization of the instrument originally developed for the *Catalog*. It collects the following information:

- Dataset Identification
  - dataset name and acronym
  - dataset description

- Information on Parent Data Bases
  - parent data base name and acronym
  - identification of datasets with data base
  - information on link variables between datasets

- Technical Dataset Information
  - name of dataset manager and responsible ARI research unit
  - dataset storage location, hardware, software, and device
  - name of dataset library (if applicable) and member (if applicable)
  - dataset catalogue status
  - dataset structure
  - name and storage location of creation programs
  - name, storage location, and storage device of associated datasets (e.g., format libraries), if applicable
  - dataset format, record length, and block size, if applicable
  - number of records and variables
  - compression status
  - accessibility of dataset at storage location

- Dataset Access
  - dataset access authorization status
  - information on obtaining access authorization

- Dataset Development and Maintenance
    - data sources
    - availability and location of dataset documentation, such as record lay-outs, codebooks, descriptions of dataset development and maintenance, and dataset contents
    - availability and location of research materials, such as a research plan, sampling plan, and data collection instruments
    - future maintenance/update plans

- Dataset Description
    - data collection methods (e.g., survey, interview)
    - types of data contained in dataset (e.g., training, performance)
    - data collection cycle, if applicable
    - dataset sort order
    - encryption status of unique identifiers, if applicable
    - relevant key words
    - assessment of data quality, including biases, consistency over time, missing data, errors, and constraints
    - sample representativeness, if applicable

- Dataset Usage
    - current useage and type of useage status
    - assessment of the extent of current dataset useage
    - assessment of the extent of future ARI dataset useage
    - assessment of the extent of future DOD-wide dataset useage

- Dataset Research Contributions
    - assessment of contributions dataset makes to current and future ARI research
    - assessment of contributions dataset makes to current and future DOD-wide research
    - assessment of contributions dataset makes to current and future research in the research community in general

- Research Projects Based on Data from Dataset
    - project name
    - project purpose
    - project sponsor
    - instrument clearance number
    - principal investigator and research unit
    - beginning and ending dates
    - description of the population and the sample
    - similar future projects anticipated?
    - publications (title, author, date, report number, work unit).

The purpose of this Level I documentation is to provide on-line technical and practical information for all datasets and allow researchers to have an opportunity to review them quickly and efficiently. Level I documentation is recommended for all extant ARI datasets so that potential future users can quickly locate information that is useful for their research. This type of documentation could prevent duplicate data collection efforts and also provide valuable sources of information for the research community in general. Even if a dataset is currently considered unimportant and not useful, it may become more valuable in the future as research priorities change. Consequently, baseline documentation is justified for all datasets. As stated above, Level I documentation will be available both in hard copy and on-line as an ACCESS file.

Level II documentation. In addition to Level I documentation, Level II documentation collects extant information that was produced as part of the original research project involving these datasets. Because of resource limitations, no attempt will be made to generate new documentation or alter the form of the existing information.

The purpose of Level II documentation is to provide supplemental information that is too complex or lengthy to be to be included in Level I documentation. By including relevant materials produced during the course of the research effort that generated the data, researchers interested in the dataset will have additional information necessary to evaluate the relevance of the dataset to their own work.

The types of information included in Level II will depend on two factors: (1) the nature of the materials produced at the time the dataset was generated, and (2) the current availability of those materials. The following is a list of the types of information that might be collected as part of the Level II documentation process:

- Research Design
- Sampling Plan and Procedures
- Data Collection Plan
- Data Collection Instruments
- Variable List
- Data Editing Procedures
- Codebook or Contents of Dataset(s)
- Data Analysis Report(s)
- Research Report(s)/Publication(s)
- Final Report

Every reasonable attempt will be made to include as many relevant extant documents as possible. To do this, the dataset POC will be given a form which lists all the types of materials that may be collected for Level II documentation and will be asked to indicate whether the material is available; this Level II Documentation Collection Form is presented in Appendix D. The dataset POC will be asked to provide all available and relevant materials produced in the original research effort for Level II documentation. The material collected will then be reviewed to determine its relevance. In no case will extracts of extant documents be made and compiled into a new document; all hard-copy research materials/publications collected will be archived in

their original hard-copy format. Electronic files will be collected and archived as they are provided to us. The goal of this level of documentation is to assemble as much information as possible on each dataset, short of generating new documents. Researchers will have more information to use in evaluating the relevance of extant datasets, but the documentation will not be consistent nor standardized across Level II datasets.

Baseline documentation for Level II datasets will be available in electronic form as well as in hard copy. ARI research and technical reports are available on-line in its Document Archival and Retrieval System (DARS); baseline documentation will include a DARS report number, when available, to facilitate the retrieval of technical reports in DARS.

Level III documentation. Level III documentation will provide additional information to that provided in Levels I and II, including variable descriptions, the distribution of values for each variable, how the variables are related to each other, and other information necessary to help new users access the existing dataset efficiently and effectively. At this level, standardized user's guides will ensure consistency of information across all Level III datasets. It is important to note that we may be unable to develop Level III documentation if key information (e.g., record layout for flat files) is not available. The overall goal of Level III documentation is to provide researchers with the information they need to use, as well as to evaluate, each dataset.

The hard copy Level III user's guide will consist of three sections:

1. Baseline documentation

2. Dataset and data element codebook

3. Available research materials

The information presented in Section One of the user's guide (baseline documentation) will enable a new user to quickly determine whether a dataset is of any interest. The second section, the nuts and bolts of the user's guide, contains the types of information typically contained in codebooks, such as variable and file attributes. Based on a review of the literature, existing documentation (Young & Kiel, 1996), and our existing knowledge of MPR datasets, we have identified three key components of codebooks that will be presented in Section Two of the user's guide:

(1) File Descriptions and Data Element List:
- Record Layout for ASCII/EBCDIC files
- CONTENTS Listing for SAS System Files
- SYSFILE INFO Listing for SPSS System Files

(2) Data Element Descriptions
- Table Contents
- Descriptive Statistics
  - Description of Formats/Value Labels
  - Frequencies, Means

(3) SAS Program Code for Formats

The first section of the codebook provides a record layout for ASCII/EBCDIC flat files and contents listing for SAS or SPSS system files. A record layout for flat files should contain the following items:

- Variable description/name
- Variable type: numeric, character, or alpha-numeric
- Variable length: number of bytes
- Variable beginning and ending column positions
- Valid codes or specific range of acceptable values
- Descriptive/Comments: specify meaning of each acceptable code, the origin of variable, or other related information
- Data Source: indicating the data collection instrument from which the variable was collected, where appropriate.

Procedures exist in both SAS and SPSS to list dataset and variable attributes. For example, the SAS CONTENTS procedure produces dataset attributes, such as the name of the SAS file, the total number of observations, and the number of variables, as well as an alphabetical list of variables, their type (either character or numeric), their corresponding length (in SAS length format), their corresponding position in the SAS file, the name of the SAS format associated with the variable, and the label associated with each variable, if applicable. The SPSS system provides comparable information in its SYSFILE INFO procedure.

The second section of the codebook, the presentation of data element descriptive statistics, is essential. The actual tabulations will be preceded by a table of contents that lists the variable name and the page number for the descriptive statistics for that variable. The descriptive statistics will include frequency distributions for discrete (nominal and ordinal level) variables and calculations of means, standard deviations, and value ranges for continuous (interval level) variables. Formatted values of variables will be presented in the frequencies, where available. In addition, a description of the variable format, where applicable, will be provided with the frequency distributions of discrete variables, so that the user can easily determine both coded and formatted values. Whenever possible, variables that link two or more datasets will be identified.

For SAS and SPSS files, descriptive statistics will be produced using the software in which the data are stored. For flat files or files generated by software other than SAS or SPSS, descriptive statistics will be produced using SAS. Since we will have to create SAS datasets to produce descriptive statistics for those datasets, we can produce SAS versions of these datasets for archiving upon ARI's request. If SAS versions are requested, we will include a SAS CONTENTS in section one of the codebook. The output containing these statistics will be modified using a PC word processor, such as Microsoft WORD or WordPerfect, and incorporated into the remainder of the codebook.

The third section of the codebook provides information on variable formats or value labels. We anticipate that many files will not contain variables with assigned formats; Level III documentation will not create formats in these cases. SAS and SPSS files differ significantly in how variable formats are stored. SPSS stores the information needed to produce formats as part of the system file. Because of this, whenever formats are assigned to readable SPSS files, we anticipate being able to use and produce a description of them.

Unlike SPSS files, SAS files write variable formats in a library or catalog that is separate from the dataset itself. We anticipate encountering a number of SAS datasets for which format assignments are recorded in the data, but format libraries/catalogs are lost or otherwise unavailable. Unless special options are used, SAS will produce errors if format libraries/catalogs are not made available to a job that is using a SAS dataset that calls for those formats. To avoid problems to future users, we will strip the format assignments from the dataset in these cases and we will not create new format libraries. For those cases where SAS format libraries/catalogs are available and readable, we will produce a copy of the program that, when executed, will generate the necessary SAS format library.[3]

Finally, section three of the user's guide will contain available materials related to the research effort from which the dataset originated, such as the research plan, the data management plan, the sampling plan, and project reports (i.e., Level II documentation).

The first two sections of the user's guide - the baseline documentation and the codebook - will be available in electronic and hard copy form. ARI research and technical reports contained in the third section of the user's guide are available on-line in DARS; baseline documentation will include a DARS report number, when available, to facilitate the retrieval of technical reports in DARS. Other research materials in section three that were originally available in hard copy form only will not be automated. Electronic files will be archived as they are presented to us.

[3] By providing a copy of the SAS program that will generate the required formats - both as hard copy and as an ASCII file stored with the dataset - the user simply needs to execute the program to create the format library. Another alternative is to save these libraries as transport datasets, archive them, and instruct the future user how to use them to create the library. We feel that the first approach is more user-friendly.

Confidentiality

While there is universal agreement in the social sciences that datasets should be well documented, there is also universal agreement that information should not be traceable to specific individuals when there are issues of confidentiality (as is usually the case). The issue of confidentiality may also exist when the unit of analysis in a dataset is not a person, but an aggregation of individuals that form an entity, such as a combat unit.

There are two approaches in dealing with the issue of confidentiality of individuals or entities. First, unique identifiers such as Social Security Number (SSN) can be stripped from the data or encrypted. The encryption approach was used in the Project A/Career Force data base. Second, access to the data containing unique identifiers can be limited to authorized users only. The Defense Manpower Data Center (DMDC) limits access to its data (which often contain unique identifiers) by limiting access to its computer facilities. Although we originally planned to encrypt unique identifiers in the course of this project, we were subsequently instructed by ARI not to do so, since ARI plans to house the archived data in a non-lending library. In producing archive datasets, ARI has also asked us to de-encrypt identifiers in those datasets where an identifier(s) was originally encrypted, as long as the encryption formula is available. At this time, ARI plans to keep the archived data in a locked safe with limited access to its contents.

Archive Media Evaluation

Datasets developed by or for ARI currently exist in many locations and are stored on various media. We began the task of evaluating various archive media by a review of the literature on the current technology of recording media.

In general, there are three criteria by which to judge various storage media (Texas State Technical College Waco, 1994):

1. Storage Capacity - The storage capacity of a recording media is usually measured in megabytes (MB).

2. Transfer Rate - This is the number of bytes that can be transferred from the media in a set amount of time and is usually measured as kilobytes per second (KB/sec) or megabytes per second (MB/sec).

3. Access Time - Access time is comprised of two measures: seek time and latency. Seek time measures how long the media takes to locate the data and latency is the amount of time before data is ready to be read. It is usually measured in milliseconds (ms).

The literature typically provides evaluations of the characteristics of five commonly available storage media: floppy disks, hard drives, cartridge tapes, zip disks, and CD-ROMs. In general, higher storage capacities and transfer rates and shorter access times are preferred.

Personal computers (PCs) have become almost universally utilized in both business and personal applications. In research applications, more and more data processing and analysis tasks are being performed on a PC platform using PC/SAS or SPSS/PC in DOS, Windows or Macintosh environments rather than on mainframe computers. Our evaluation of archive media assumes that the computing platform of choice at ARI is the PC. Of the storage media evaluated here, floppy disks, zip disks, and CD-ROM are unique to the PC environment; cartridge tapes and permanent drives may apply to both the PC and mainframe environments. Table 2 presents the status of these five archive media on the three criteria outlined above. In addition, Table 2 summarizes the advantages and disadvantages of each of media based on such additional archive media characteristics as portability, usability, ease of storage, and cost efficiency.

Although floppy disks are inexpensive and portable, their low storage capacity is a serious limitation, particularly when storing datasets of considerable size. Personal computer hard drives possess many advantages, including fast transfer rates, low access times and typically large storage capacity. Unfortunately, hard drives[4] do not satisfy the portability requirement since the recording media is mounted within the case itself.

Cartridge tapes have typically been used in mainframe computer facilities, although tapes for PC drives do exist. One advantage of cartridge tapes is that they can hold large amounts of data. A disadvantage is that cartridge tapes can be over-written. Furthermore, without proper care and handling, cartridge tapes and other magnetic tapes are subjected to chemical, mechanical, and/or magnetic failure (Williamson, 1991). Many mainframe computer facilities, such as the National Institutes of Health (NIH), keep thousands of 3480/3490 cartridge tapes in their tape library. Although many of the ARI extant datasets are currently stored on 3480 cartridge tapes at NIH, ARI staff and contractors can only read those tapes at NIH or other MVS mainframe facilities. While these 3480/3490 cartridge tapes are portable in the strictest sense of the word, the availability of drives that read them and the ease with which a researcher can use the facilities that operate those drive may be a serious disadvantage.

Zip disks are written and read by Zip drives and are another portable device for data storage. Zip disks have a storage capacity from 100 MB to 1.15 GB and transfer rate that is similar to hard drives. Zip disks are a magnetic recording media. Therefore, problems resulting from improper handling and storage of magnetic media also apply to zip disks. Currently, a potentially serious disadvantage to Zip technology is that the disks and drives are not as commonly used as other recording devices.

CD-ROM brings together the best features of the other recording media evaluated here. CD-ROM drives on PCs are becoming increasingly popular. Like floppy and Zip disks, CD-ROMs are removable, relatively inexpensive, and easy to use. Unlike floppy disks, CD-ROMs

[4] Although not appropriate for this archive effort, a hard drive on a Web Site Server would be a viable media for storing data that could be accessed through the World Wide Web.

Table 2. Comparisons of Five Data Storage Media

Characteristics	Floppy Disks	Hard Drives	Cartridge Tapes	Zip Disks	CD ROMs
Storage Capacity	360 KB to 1.44 MB	80 MB to 2 GB	250 MB to 1 GB	100 MB to 1.5 GB	680 MB
Transfer Rate	150 KB/sec	1-4 MB/sec	1-4 MB/sec	1-4 MB/sec	150, 300, 450, 600 KB/sec
Access Time	500 ms	9-30 ms	15-30 ms	15-30 ms	150-350 ms
Advantages	- Cheap - Transportable	- Fast transfer rates - Low access times - Large storage capacity	- Large storage capacity - High transfer rates - Transportable	- Fast transfer rates - Low access time - Large storage capacity - Transportable	- Large storage capacity - Transportable - Widely in use - Relatively cheap to produce
Disadvantages	- Transfer rate low - Access time high - Low storage capacity	- Not transportable	- Not widely used - Need special tape drives	- Not widely used - Need special drives	- Need CD-Recordable technology to create

provide far more data storage capacity. Recording media longevity and, as a consequence, data durability, does not seem to differ substantially between CD-ROM and floppy disks[5]. However, unlike data on floppy and Zip disks and tape cartridges, data recorded on CD-ROM cannot be over-written. Eliminating the possibility of accidental data over-writing is clear advantage of the CD-ROM over other archive media. In preparing guidelines for DoD produced CD-ROM products, the Defense Information Systems Agency (DISA) states that " the use of CD-ROM to store and disseminate information is not only becoming reality, but is being implemented throughout DOD as a means of reducing paper/magnetic media/microform distribution and attendant costs (DISA, 1995, p.ii)." This document suggests that the use of CD-ROM for data archive and data dissemination is greatly beneficial to the DOD community. The document further provides many guidelines for the development and distribution of CD-ROMs. It further suggests that ISO 9660 format be used for all CD-ROMs developed within DoD. ISO 9660 is the international standard which defines the file structure for putting computer files on compact discs. It is also widely used in commercial applications. The DISA document also provides general guidelines for the physical properties of CD-ROM, disc labeling, data classification, handling caveats, and documenting disc contents. These guidelines have been further adapted for ARI use (Speight, 1996).

Overall, our evaluation suggests that the use of CD-ROM for data archive is the data recording media of choice for PC based applications. For those datasets that will continue to be used on a mainframe, 3480 cartridges may serve as the archive media as well, depending on the dataset POC's recommendation.

Conclusions

Documentation standards of extant datasets must be responsive to differing assessments of the relative cost and benefit of additional documentation, as well as to technical characteristics of the data. Therefore, we have developed a system of standards that is responsive to different documentation requirements. The three-tier documentation system groups datasets into requirement categories based on the assessments of ARI staff most familiar with the data. In addition, the system is responsive to practical considerations, such dataset availability, and technical characteristics, such as file structure. Through a review of industry standards and our own extensive experience with ARI MPR data, we have established documentation standards for each of these data documentation requirement levels. In addition, we have evaluated different recording media and identified the archive media of choice. This system of documentation requirements and corresponding documentation/archive standards will ensure that, within practical cost-tradeoff parameters, future users will have the necessary information to assess and utilize extant ARI MPR datasets.

---

[5] Based on a conversation with and written communication from Paul Simon, Technical Service, Imation Data Storage Division, 3M, data stored on floppy disks (3 1/2") can last up to 100 years or more, if stored and accessed appropriately. Data stored on CD-R (CD-Recordable) discs will also last 100 years or more.

References

American Psychological Association (1992). Ethical principals of psychologists and code of conduct. Washington, DC: Author.

American Statistical Association (1993). Ethical guidelines for statistical practice. Alexandria, VA: Author.

The Defense Information Systems Agency (1996, December). Department of Defense Handbook: DoD-Produced CD-ROM Products (MIL-HDBK-9770). Washington, DC: Author.

Defense Manpower Data Center (1996). DMDC data dictionary. Arlington, VA: Author.

Department of Defense, Assistant Secretary of Defense for Command, Control, Communications, and Intelligence (1994). Data administration procedures. (DTIC ADA282 718). Washington, DC: Author.

Department of Defense, Office of the Under Secretary of Defense (Personnel and Readiness) (1996). DOD personnel data standardization administration: Draft DOD personnel data model, version 8.0. Arlington, VA: Author.

Fienberg, S. E. (1994). Sharing statistical data in the biomedical and health sciences: Ethical, institutional, legal, and professional dimensions. Annual Review of Public Health, 15.

Holway, J., Tao, F., Ramsey, L., Payne, R., & Haupt, A. (1992). Catalog and assessment of the Manpower and Personnel Research Division data bases. Arlington, VA: Fu Associates.

Inter-university Consortium for Political and Social Research (1996). Guide to social science data preparation and archiving. Ann Arbor, MI: Author.

Peterson, J., & Colella, U. (1991). Depositing data with the Data Resources Program of the National Institute of Justice: A handbook. Washington, DC: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Roistacher, R. C. (1980). A style manual for machine-readable data files and their documentation (Report number SD-T-3, NCJ-62766). Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Speight, N. (1996, October). Guidelines for Information Products on Compact Disk (CD). Memorandum for See Distribution. Alexandria, VA: Army Research Institute.

SPSS (1993). SPSS for Windows, base system user's guide, release 6.0. Chicago, Il: Author.

Statistical Analysis System (1990). SAS procedures guide, version 6. Cary, NC: Author

Texas State Technical College, Waco (1994). Developing CD-ROMs: Pitfalls and Detours on the road to the Digital Village. From a presentation given at the League of Innovation in Community Colleges.

U.S. Army Research Institute (1995). Sample of ARI Databases. Unpublished manuscript. Alexandria, VA: Author.

U.S. Department of Justice, Office of Justice Programs, National Institute of Justice (1996). Data Resources Program of the National Institute of Justice. Ann Arbor, MI: Author

Williamson, M.P. (1991). The 3480 type tape cartridge: potential data storage risks, and care and handling procedures to minimize risk. Gaithersburg, MD: National Institute of Standards and Technology.

Wise, L., & Wang, M. (1983). Development and validation of Army selection and classification measures, Project A: Longitudinal research database plan (Final Report 83-12). Alexandria, VA: U.S. Army Research Institute.

Young, W., & Keil, C. T. (1996). Longitudinal validation experimental predictor battery: SASB5ALLV01 codebook (draft). Alexandria, VA: U.S. Army Research Institute.

APPENDIX A

ARI Staff Survey

ARI Staff Survey

Instructions for Survey Completion

This survey contains a list of ARI datasets. The purpose of the survey is to identify individuals who are familiar with the critical datasets used by ARI and to quantify the usage and importance of these datasets. Please read each dataset description and answer the questions that follow the description. Specific answers we are interested in are:

NOT APPLICABLE: If you have not used and do not plan to use a dataset in the future, mark this line with an X immediately after the colon (NOT APPLICABLE:X) and skip the other questions for this dataset.

OWNERSHIP: If you directed or performed the initial development of the dataset, select Principal Investigator; do not select Primary User. If you currently maintain or control the dataset or you are one of the heaviest users of the dataset, select Primary User. The intent of the Ownership line is to identify individuals who may be able to provide information on the construction or content of the dataset. Other users may skip the Ownership question and proceed to the next question.

FREQUENCY OF CURRENT USE: Select the appropriate level of use that describes your usage of the dataset within the last two years.

FREQUENCY OF FUTURE USE: Select the appropriate level of use based on how often you expect or intend to utilize the dataset in the future.

IMPORTANCE: Select the appropriate response based on what you believe to be the relative importance of the dataset to your current and/or future work.

After you have reviewed all the descriptions, if you find that you use datasets that were developed by ARI and are not included on this list, please add them in the spaces provided at the end of the list. Complete all information for the dataset(s) you add. Increment the dataset number for each dataset you add. Use as many lines as you need to enter full descriptions of the datasets.

Note that there is a distinction drawn between datasets and databases. For this survey, a dataset may be one of several files that are derived from a parent database. For example, there is a file containing ROTC data (the ROTC Dataset) among the files that make up the Officer Longitudinal Research Data Base (OLRDB). The OLRDB would be represented on each dataset description as the Parent Data Base.

This survey is intended to solicit your requirements for datasets, with notations about the relationship to a parent database, if relevant. Please provide a separate entry and description for each additional dataset you utilize, even if more than one dataset belongs to the same parent data base.

DIRECTIONS:

Please read the description for each dataset. Answer each of the questions below the dataset description by putting the number associated with the selected response alternative immediately after the RESPONSE: for each question. For example, if you use the dataset eight times per year, your entry should look like this:

QUESTION2: Frequency of Current Use
CHOICES: 1=Never Use, 2=One to Four Times/Yr, 3=Five+ Times/Yr
RESPONSE2:3

If you have any questions or problems completing the form, please call Dianne Driessen at Fu Associates, Ltd. at (703) 243-2992.

When you are finished with the survey, please forward it through the Internet to Dianne Driessen at driessen@nerdvana.fu.com. Thank you for taking the time to provide this information.

****
DATASET SURVEY LIST
****

DATASET NUMBER: *(a data set number is provided for each dataset)*
DATASET NAME: *(dataset name is provided)*
PARENT DATABASE: *(parent data base information is provided, if applicable)*
DATASET DESCRIPTION: *(a description of each dataset is provided)*

NOT APPLICABLE:

QUESTION1: Ownership
CHOICES: 1=Principal Investigator, 2=Primary User
RESPONSE1:

QUESTION2: Frequency of Current Use
CHOICES: 1=Never Use, 2=One to Four Times/Yr, 3=Five+ Times/Yr
RESPONSE2:

QUESTION3: Frequency of Future Use
CHOICES: 1=Never Use, 2=One to Four Times/Yr, 3=Five+ Times/Yr
RESPONSE3:

QUESTION4: Importance
CHOICES: 1=Not Important, 2=Somewhat Important, 3=Very Important
RESPONSE4

APPENDIX B

Recommended Archive Documentation Level Questionnaire

Recommended Archive Documentation Level

(Dataset Name)
(POC)

You are the Point of Contact (POC) for the above mentioned dataset. As you may know, ARI has engaged the services of HumRRO and Fu Associates to document and archive its Manpower and Personnel datasets. As POC, we are interested in your opinion on the level of documentation this dataset requires. We have delineated three levels of documentation and provide a description of each below. Please indicate which **one** of the three levels of documentation your dataset requires on the response line provided below.

DOCUMENTATION LEVELS

Level I Baseline Documentation

This is the lowest level of documentation a dataset can have. You may already have provided baseline information for your dataset to Fu staff members. Baseline documentation contains such technical and descriptive information as:

parent database information, if applicable
dataset storage location, computer, operating system, storage device
dataset configuration (e.g., flat or system file)
dataset format
required associated files, such as format libraries or indexes
names and locations of creation programs
dataset access
description of data sources
available documentation
update schedule, if applicable
data collection method and type
data sort order
encryption status of unique identifiers
general data quality, overall accuracy, completeness; sample representativeness, known biases, and dataset constraints, if applicable
dataset usage and contributions to research
description of research project based on data from the dataset

Level II *Intermediate Documentation*

This documentation level includes the baseline documentation described in Level I above, as well as a collection of associated available and relevant research materials. These research materials may include the following:

research design
sampling plan and procedures
data collection plan
data collection instruments
variable lists/dataset contents
data editing procedures
codebooks/user's guides
technical reports
other

Level III *Advanced Documentation*

This documentation level includes all the elements of the intermediate-level documentation (LEVEL II above) as well as a detailed codebook. The codebook will consist of the following:

File descriptions and data element list (e.g., a record layout for flat files, a CONTENTS for SAS files, and SYSFILE INFO for SPSS files)

Data element descriptions
- descriptive statistics
- format/value label descriptions

Program code for SAS format library, if applicable

RESPONSE

It is my opinion that the above mentioned dataset requires Level (I, II, or III) documentation.

Signature Date

APPENDIX C

ARI Research Dataset Questionnaire

ARI RESEARCH DATASET QUESTIONNAIRE

ARI is in the process of systematically documenting and archiving ARI research datasets. The purpose of this questionnaire is to collect dataset information from ARI staff most knowledgeable about the data. The information collected will be used in archiving the datasets and generating further documentation if applicable. The information will also be entered into the ARI Dataset Documentation Database.

IMPORTANT

Most of the items in this form are self explanatory. However, a critical first step for the POC is to review the definitions of "datasets" and "databases" below and determine the target "dataset" to be described in each questionnaire. Research datasets commonly belong to a group of related datasets. The Dataset Classification Table on the following page describes four typical arrangements by which datasets are organized and stored. The table also provides examples of ARI datasets that represent each arrangement and indicates how each arrangement should be described in this questionnaire. Before you begin to fill a form, please determine which of these arrangements applies to the dataset you are describing and how many separate forms need to be completed for that dataset.

If you need clarification on the number of questionnaires to describe "a dataset" please contact Donna Peck at 703-243-2992.

DATASET CLASSIFICATION TABLE

Parent Database

For this questionnaire, a database refers to a group of datasets, each containing different types of data, but collectively belonging to, or derived from, the same research project. A parent database is not a physical data file. Each member dataset of a database is documented by a separate ARI Research Dataset Questionnaire.

HOW TO COMPLETE: The focus of, or the unit of analysis for, this questionnaire is a dataset. If a dataset belongs to a parent database, complete a separate questionnaire for each dataset in a parent database and indicate the name of the parent database under Section A.

Each of the datasets in a parent database may represent different dataset organizations as described below in this table.

Example:

The Officer Longitudinal Research Data Base (OLRDB) includes: the Core Dataset with personnel data from the annual Officer Master File; the ROTC Dataset with ROTC precommissioning training data; AIMS Dataset with basic and advanced training data; and several other datasets containing different types of data.

Multinational Force and Observers - Sinai (MFO) database consists of the Family/Finance dataset, Cohesion/Leadership datasets, Job Knowledge datasets, and others.

Dataset With Multiple Physical Data Files

all containing a similar set of variables and supporting the same research project or objective

Multiple data files contain virtually the same number and type of variables, have the same file characteristics are stored in the same medium at the same location. The files have different file names, and may differ in number of records, year and/or location of data collection, and the name of program used to create system files.

HOW TO COMPLETE: One questionnaire is completed for all data files in the dataset. Provide a multiple listing of all component data files with file specifications and characteristics that differ among the files (e.g., file names, number of variables, number of records, year of data collection) using a table attached at the end of the questionnaire (one table for listing SAS or SPSS system files, another table for listing flat files).

Example:

EMF Dataset consisting of 140 physical files of data extracted from the Enlisted Master File every quarter.

Building the Career Force/Longitudinal Validation I, Batch A MOS dataset consists of 10 SAS system files containing data collected from 20 MOSs.

Datasets That Support the Same Project and/or Research Objectives

but differ substantially in contents and/or file characteristics

Datasets may represent data from a survey that changed significantly in survey format, contents, sample group, or phases of a research project (e.g., at enlistment vs. during the second tour).

HOW TO COMPLETE: Complete a separate questionnaire for each dataset even if the responses to many items may be similar (e.g., purpose of the research project, storage medium and location, physical file characteristics, dataset documentation, and extent of current and future use of the data).

Example:

Bosnia Pre-Deployment, During Deployment, and Post-Deployment Soldier Surveys.

Multiple MFO - Sinai Cohesion/Leadership Datasets differing in the time and location of administration, samples and survey items.

Datasets containing data from the semi-annual Sample Survey of Military Personnel; each survey and the resulting dataset address different issues.

Single, Unitary Dataset

One physical data file containing all data collected for a specific research project. It may be the end product of merging originally separate files.

HOW TO COMPLETE: Complete one questionnaire for each unitary dataset.

Example:

MFO - Sinai Family/Finance Dataset
Special Forces Concurrent Validation Dataset
ABLE Coaching Dataset.

ARI RESEARCH DATASET QUESTIONNAIRE

Name of Respondent: _____

Research Unit: _____ Date of Questionnaire: ____/____/____

Dataset Name: _____

Dataset Acronym: _____

Brief Dataset Description (Include discussion of purpose of the research and major areas of investigation period covered by data, and nature of data):

_____

_____

_____

_____

_____

_____

NOTE: If the dataset contains many years of longitudinal data or many sub-files, some of which may be waived from complete archiving, which files should be documented and archived in this project (specify by years covered by data, contents of sub-files, etc.)?

_____

_____

_____

_____

_____

A. Parent Data Base

If the dataset belongs to a Parent Database, complete this section, otherwise proceed to Section B.

1. Parent Data Base Name (If applicable): _____
2. Parent Data Base Acronym: _____
3. How many datasets are included in the data base? _____
4. Can each dataset be used separately for analysis? ___ Yes ___ No
5. Can the datasets within the data base be linked? ___ Yes ___ No
6. What variable(s) are used to link the datasets (e.g., SSN)?

   _____

B. Technical Dataset Information

The purpose of this section is to gather the information necessary to locate and read the physical dataset for archival. If the dataset contains multiple physical data files, omit this section and complete either Table 1 for SAS and SPSS files or Table 2 for raw/flat file.

1. Dataset management:
   a) ARI Research Unit responsible: _____
   b) Dataset manager: _____

2. Currently, where is the dataset stored/maintained?
   ___ ARI VAX
   ___ ARI PC Which ARI PC (e.g., location, user) ? _____
   ___ NIH
   ___ Other/specify: _____

3. What type of computer is used to store the dataset (e.g., IBM mainframe, DEC, SUN, PC, Macintosh)?
   _____

4. What type of operating system is used to store/maintain the dataset (e.g., MVS, VMS, Unix, DOS)?
   _____

5. What is the operating system, permanent dataset name (i.e., fully-qualified dataset name, DSN, filename)? Include the operating system dataset library name, if applicable.
   _____

6. Is the dataset catalogued (if applicable)? ___ Yes ___ No

7. What type of device is used to store the dataset (e.g., magnetic tape, 3480 cartridge, disk pack, floppy disk, CD ROM)?
   ___ Magnetic tape
   Volume/serial numbers: _____
   Standard labeled/nonlabeled: _____
   BPI (800, 1600, 6250, or other): _____
   ___ 3480 cartridges
   Volume/serial numbers: _____
   Standard labeled/nonlabeled: _____
   ___ Online disk: Volume/serial numbers: _____
   ___ Online dedicated disk: Volume/serial numbers: _____

___ Floppy disk: Density: _____ Number of disks: _____

___ CD ROM: Number of CDs: _____

___ Other (describe): _____

8. Is the dataset a structured, system file format dataset (e.g., SAS, SPSS, VSAM, RDBMS)?

___ Yes ___ No (Go to Question B.9.)

a) If yes, what format (e.g., SAS, SPSS, VSAM, Oracle)? _____

b) What version (if applicable)? _____

c) Internal dataset or member name (if applicable)? _____

d) Table name (if applicable)? _____

e) Number of variables or columns? _____

f) Number of observations/records or rows? _____

g) Name(s) of the program(s) to create the dataset? _____

_____

h) Storage location of the program(s) to create the dataset? _____

9. Are there additional required datasets associated with this dataset (e.g., format libraries, indexes)?

___ Yes ___ No (Go to Question B.10.)

a) If yes, what additional datasets are required?

_____

b) What is the permanent, operating system dataset name of the additional required dataset(s)?

_____

c) Currently, where is the additional required dataset(s) stored/maintained?

___ ARI VAX

___ ARI PC Which ARI PC (e.g., location, user) ? _____

___ NIH

___ Other/specify: _____

d) What type of computer is used to store the additional required dataset(s) (e.g., IBM mainframe, DEC, SUN, PC, Macintosh)?

_____

e) What type of operating system is used to store/maintain the additional required dataset(s) (e.g., MVS, VMS, Unix, DOS)?

_____

f) Name(s) of the program(s) to create the additional required dataset(s)?

_____

g) Storage location of the program(s) to create the additional required dataset(s)?

_____

h) What type of device is used to store the additional required dataset(s) (e.g., magnetic tape, 3480 cartridge, disk pack, floppy disk, CD ROM)?

___ Magnetic tape

Volume/serial numbers: _____

Standard labeled/nonlabeled: _____

BPI (800, 1600, 6250, or other): _____

___ 3480 cartridges

Volume/serial numbers: _____

Standard labeled/nonlabeled: _____

___ Online disk: Volume/serial numbers: _____

___ Online dedicated disk: Volume/serial numbers: _____

___ Floppy disk: Density: _____ Number of disks: _____

___ CD ROM: Number of CDs: _____

___ Other (describe): _____

10. If the dataset is in raw/flat file format:

a) Record format (e.g., fixed block, variable block)? _____

b) Record length (e.g., LRECL)? _____

c) Block size? _____

d) ASCII or EBCDIC? _____

e) Number of variables? _____

f) Number of observations/records: _____

11. Is the dataset compressed?

___ Yes ___ No (Go to Question B.12.)

a) If yes, check the appropriate format:

___ UNIX TAR file ___ UNIX compressed file

___ pkZip file ___ GZip file

___ Other compressed file (please specify): _____

12. Is this dataset readily accessible at the location and on the medium described above?

___ Yes ___ No

Comments: _____

C. Dataset Access

Provide information necessary to ensure that the archival team can access the dataset and understand any restrictions necessary for use of the archived data.

1. For data stored at NIH, is RACF (Resource Access Control Facility) authorization required?

   ___ Yes ___ No (Go to Question C.2.)

   a) If yes, how can a user obtain the RACF authorization?

   _____

2. Is the access to this dataset or certain variables in the dataset restricted to general users except through the Freedom of Information Act (FOIA) request process (e.g., "close-hold" dataset and/or variables whose release is controlled by a sponsor)?

   ___ Yes ___ No (Go to Question C.3.)

   a) If yes, does the restriction apply to the entire dataset or certain variables in the dataset?

   _____

   _____

   b) If only selected variables of the dataset are restricted, can the remaining variables be accessed by general users and how?

   _____

   _____

   _____

   _____

3. Is any other special authorization required to access the dataset?

   ___ Yes ___ No (Go to Question C.4.)

   a) If yes, what authorization is required?

   _____

   b) How can a user obtain the authorization? _____

4. What is the office name or person to contact to obtain further information on how to access the dataset?

   _____

D. Dataset Development and Maintenance

This section is intended to provide future users with the information necessary to understand the construction and coding of dataset variables and the schedule for future update, if applicable.

1. What are the sources of data for this dataset (e.g., name of survey, administrative database)?

_____

_____

2. Is documentation of the dataset development process available?

___ Yes ___ No (Go to Question D.3.)

a) If yes, is the dataset development documentation available online or in hardcopy?

___ Online ___ Hardcopy

a.1) Online location: _____

a.2) Hardcopy location: _____

b) Does the documentation include editing/ manipulation specifications?

___ Yes ___ No

3. Is dataset codebook/data dictionary (variable descriptions) available?

___ Yes ___ No (Go to Question D.4.)

If yes, does the codebook/data dictionary contain the following information?

a.1) Variable description

___ Yes ___ No

a.2) Variable code/value description (e.g., 1=male; 2=female) ___ Yes ___ No

a.3) Description of constructed variables ___ Yes ___ No

a.4) Variable descriptive statistics ___ Yes ___ No

b) If yes, is the codebook available online or in hardcopy?

___ Online ___ Hardcopy

b.1) Online location: _____

b.2) Hardcopy location: _____

4. If the dataset is in raw/flat file format, is the dataset's record layout available?

___ Yes ___ No (Go to Question D.5.)

a) If yes, is the record layout available online or in hardcopy?

___ Online ___ Hardcopy

a.1) Online location: _____

a.2) Hardcopy location: _____

5. Is any other dataset documentation available (e.g., SAS contents listing, SAS format library listing, SPSS file information)?

___ Yes ___ No (Go to Question D.6.)

a) If yes, please describe the type of other dataset documentation.

_____

_____

b) If yes, is the additional documentation available online or in hardcopy?

___ Online ___ Hardcopy

b.1) Online location: _____

b.2) Hardcopy location: _____

6. Are materials such as research plan, sampling plan, data collection instruments, and others that are relevant to the dataset available?

___ Yes ___ No (Go to Question D.7.)

a) If yes, are these materials available online or in hardcopy?

___ Online ___ Hardcopy

a.1) Online location: _____

a.2) Hardcopy location: _____

7. Will this dataset be updated in the future? ___ Yes ___ No (Go to Question 8)

a) If yes, planned update schedule? _____

b) What is the established update procedure (who does what when)?

_____

_____

_____

c) Is documentation of the dataset update process available? ___ Yes ___ No

c.1) If yes, is the dataset update documentation available online or in hardcopy?

___ Online ___ Hardcopy

c.2) Online location: _____

c.3) Hardcopy location: _____

8. Supporting Documentation

Please provide as much supporting documentation as possible, including but not limited to the following.

___ Research Design

___ Sampling Plan and Procedures

___ Data Collection Plan

___ Data Collection Instruments

___ Variable List

___ Data Editing Procedures

___ Codebook or Contents of Dataset(s)

___ Data Analysis Report(s)

___ Research Report(s)/Publication(s)

___ Final Report

___ Other, Please Specify _____

E. Data Description

This section describes the research methods and keywords associated with the dataset, population or sample, and data completeness and usefulness.

1. What methods were used to collect data contained in this dataset? (Check as many as apply.)

___ Survey

Indicate the PT Number (survey approval number from APSO, the U.S. Army Personnel Survey Office): _____

___ Interview

___ Performance assessment

___ Cognitive test

___ Attitude assessment

___ Aptitude/achievement tests

___ Observational

___ Copy of existing management data base

___ Extracted from existing management data base

___ Other (list) _____

2. What type(s) of data are contained in this dataset (e.g., personnel, test scores, training performance)?

_____

_____

3. If these data are collected on a periodic cycle (e.g., annually), what years are covered by this dataset?

Fiscal year: _____ Calendar year: _____

4. What is the sort order of data in the dataset? _____

_____

5. Does the dataset include social security numbers?

___ Yes___ No (Go to Question E.6.)

a) If yes, are the social security numbers encrypted? ___ Yes ___ No

a.1) If yes, is the encryption formula available? ___ Yes ___ No

a.2) If yes, what office or person can provide the encryption information?

_____

6. What keywords are relevant to the dataset?

a) Concepts, topic area keywords (e.g., recruitment, re-enlistment, leadership training, family services):

_____

_____

_____

b) Measures keywords (e.g., aptitude test): _____

_____

_____

c) Sample groups or population keywords (e.g., commissioned officers, enlisted personnel):

_____

_____

_____

d) Sponsor, related organizational units keywords (e.g.,TRADOC, DCSPER, DCSOPS, FORSCOM, NTC):

_____

_____

7. Describe the overall accuracy of data in this dataset.

_____

_____

8. Describe the overall completeness of data in this dataset. If the dataset is designed to include the entire population, do the actual data in the dataset represent the population fully?

_____

_____

9. Does the dataset have any known bias (due to sampling or data collection procedures)?

_____
_____
_____

10. If the data were collected over time, did the contents of the dataset change over time (e.g., variables, value coding system)?

_____
_____
_____

11. What are the dataset constraints, if any (e.g., size of dataset, execution speed, too few variables, inability to link with other datasets)?

_____
_____

12. Representativeness:

a) Does the dataset contain data from a sample(s) of a population or does it describe the entire universe?

___ sample ___ universe (If a universe, skip to item 12c.)

b) What is the actual, overall representativeness of the sample(s)?

Rating (circle one)

1	Poor	The sample(s)/data are not representative of, or poorly represents, relevant segments of the population
2		
3	Fair	The sample(s)/data approximate some segments of the population, but some specialized areas (e.g., MOS) may not be representative
4		
5	High	The sample(s)/data are representative for every relevant segment of the population

c) Define the universe.

_____

Comments: _____

F. Dataset Usage

1. Is anyone using the dataset now?

   ___ Yes

   ___ No When was the dataset last used? _____ (Go to Question F.4.)

2. Who is currently using data from the dataset?

   _____

3. How is the dataset being used (e.g., data updates, data analysis)?

   _____

   _____

4. To what extent is the dataset currently used (in terms of frequency of use, number of users, and/or criticality of research it supports)?

   Rating (circle one)

1	Not Extensively	Used infrequently and/or for non-critical purposes
2		
3	Moderately	Used fairly frequently and/or for somewhat critical purposes
4		
5	Very Extensively	Used very frequently and/or for highly critical purposes

Comments: _____

5. To what extent do you think the dataset will be used for ARI research in the future (in terms of frequency of use, number of users, and/or criticality of research it supports)?

   Rating (circle one)

1	Not Extensively	Used infrequently and/or for non-critical purposes
2		
3	Moderately	Used fairly frequently and/or for somewhat critical purposes
4		
5	Very Extensively	Used very frequently and/or for highly critical purposes

Comments: _____

6. To what extent do you think the dataset will be used within DOD in general in the future (in terms of frequency of use, number of users, and/or criticality of research it supports)?

Rating (circle one)

1	Not Extensively	Used infrequently and/or for non-critical purposes
2		
3	Moderately	Used fairly frequently and/or for somewhat critical purposes
4		
5	Very Extensively	Used very frequently and/or for highly critical purposes

Comments: _____

G. Dataset Research Contribution

1. Does this dataset make an important contribution to *current ARI* research?

Rating (circle one)

1	Not Important	Dataset contributes only minimally, if at all
2		
3	Moderately Important	Dataset makes a moderately important contribution
4		
5	Very Important	Dataset makes a unique and important contribution

Comments: _____

2. Do you expect that this dataset will make an important contribution to *future ARI* research?

Rating (circle one)

1	Not Important	Dataset contributes only minimally, if at all
2		
3	Moderately Important	Dataset makes a moderately important contribution
4		
5	Very Important	Dataset makes a unique and important contribution

Comments: _____

3. Do you think this dataset makes an important contribution to *current DOD-wide* research?

Rating (circle one)

1	Not Important	Dataset contributes only minimally, if at all
2		
3	Moderately Important	Dataset makes a moderately important contribution
4		
5	Very Important	Dataset makes a unique and important contribution

Comments: _____

4. Do you expect that this dataset will make an important contribution to *future DOD-wide* research?

Rating (circle one)

1	Not Important	Dataset contributes only minimally, if at all
2		
3	Moderately Important	Dataset makes a moderately important contribution
4		
5	Very Important	Dataset makes a unique and important contribution

Comments: _____

5. Does this dataset make an important contribution to current research throughout the general research community (e.g., universities, other government agencies, private research organizations)?

Rating (circle one) or indicate "N/A" for not applicable in comments below.

1	Not Important	Dataset contributes only minimally, if at all
2		
3	Moderately Important	Dataset makes a moderately important contribution
4		
5	Very Important	Dataset makes a unique and important contribution

Comments: _____

6. Do you expect that this dataset will make an important contribution to future research throughout the general research community?

Rating (circle one)

1	Not Important	Dataset contributes only minimally, if at all
2		

3	Moderately Important	Dataset makes a moderately important contribution
4		
5	Very Important	Dataset makes a unique and important contribution

Comments: _____

H. Research Project/Analysis Based On Data From The Dataset

All questions under Section I will be completed for each KEY research project or analysis based on data from the dataset. Make additional copies of this section if needed for additional projects/analyses.

Name of Respondent: _____ Research Unit: _____
Date: ____/____/____ Dataset Name: _____

1. Name of project/analysis: _____
2. Purpose of research/analysis: _____
3. Sponsor: _____
4. Instrument clearance approval numbers (e.g., OMB, USAPIC): _____
5. ARI Research Unit responsible for research/analysis: _____
6. ARI principal investigator: _____
7. Begin/end dates of research/analysis: _____
8. Description of research/analysis population:
   _____
   _____
9. Description of research/analysis sample:
   _____
   _____
10. Will similar project/research be conducted in the future?
    _____
    _____

11. Publications related to the research project or analysis. If the complete information is not available, provide sufficient information, e.g., the document title and/or the first author, to facilitate a search in the Document Archive Retrieval System (DAR).

Title: _____

_____

Author(s): _____

_____

Date: _____ Report #: _____ Work Unit #: _____

Title: _____

_____

Author(s): _____

_____

Date: _____ Report #: _____ Work Unit #: _____

Title: _____

_____

Author(s): _____

_____

Date: _____ Report #: _____ Work Unit #: _____

Title: _____

_____

Author(s): _____

_____

Date: _____ Report #: _____ Work Unit #: _____

Title: _____

_____

Author(s): _____

_____

Date: _____ Report #: _____ Work Unit #: _____

Title: _____

_____

Author(s): _____

_____

Date: _____ Report #: _____ Work Unit #: _____

Table 1

MULTIPLE SYSTEM FILES

Multiple Data Files Belonging to Dataset Name (from Page 1): _____

Type of files: _____ SPSS _____ SAS _____ Other

Location of files listed below: _____

Data compressed: _____ Yes (How? _____) _____ No

Storage Medium: _____

Are the data readily accessible? _____ Yes _____ No

ARI Research Unit Responsible: _____

Dataset Manager: _____

Type of computer used to store data: _____

Operating System: _____

Datasets catalogued?: _____ Yes _____ No

Data File Name	Number of Observations	Number of Data Fields	System File Creation Program - File Name	Number of disks and or CDs *(if applicable)*	Tape # Vol Ser No *(if applicable)*	Standard Labeled (SL) /nonlabeled (NL) *(if applicable)*	BPI Density *(if applicable)*	Characteristics Unique to this File *(if applicable)*

Table 2

MULTIPLE FLAT FILES

Multiple Data Files Belonging to Dataset Name (from Page 1): _____

Type of flat file: _____ ASCII _____ EBCDIC

Location of files listed below: _____

Data compressed: _____ Yes (How? _____) _____ No

Storage Medium: _____

Are the data readily accessible?: _____ Yes _____ No

ARI Research Unit Responsible: _____

Dataset Manager: _____

Type of computer used to store data: _____

Operating System: _____

Datasets catalogued?: _____ Yes _____ No

Data File Name	Number of Records	Number of Data Fields	Number of disks and/or CDs *(if applicable)*	Tape # Vol/ Ser No. *(if applicable)*	Standard Labeled (SL) nonlabeled (NL) *(if applicable)*	BPI Density *(if applicable)*	LRECL	Block Size	Format Fixed Block (FB) Variable Block (VB)	Characteristics Unique to this File *(if applicable)*

Appendix D
Level II Documentation Collection Form

ARI Research Dataset
Level II Documentation Collection Form

Instructions: Please put "Yes" or "No" on the left hand side of each category to indicate whether document(s) are collected for that category. If the answer is "Yes", please write down the name of the document by using the space provided on the right. If more than one document is collected for that category, please indicate total number of documents collected and then list each document separately. Attach a separate sheet of paper if necessary. If the answer is "No", please write down further explanation if provided: e.g., "In Progress", "Can't find off-hand, but will provide later", "In press, will be available in two months", "Under review", etc.

If the document is available through the Document Archive Retrieval System (DARS), please provide DARS number whenever possible. If the document is not available through DARS but an electronic copy is available, please obtain a copy on diskette.

Type of Information	*Report(s) containing the documentation*	Is this in DARS?	Electronic copy available?
_____Research Design	_____	DARS # _____	Yes No
	_____	DARS # _____	Yes No
_____Sampling Plan and Procedures	_____	DARS # _____	Yes No
	_____	DARS # _____	Yes No
_____Data Collection Plan	_____	DARS # _____	Yes No
	_____	DARS # _____	Yes No
_____Data Collection Instruments	_____	DARS # _____	Yes No
	_____	DARS # _____	Yes No
_____Variable Lists	_____	DARS # _____	Yes No
	_____		

ARI Research Dataset
Level II Documentation Collection Form

Type of Information	*Report(s) containing the documentation*	Is this in DARS?	Electronic copy available?
_____Data Editing Procedures	_____	DARS # _____	Yes No
	_____	DARS # _____	Yes No
_____Codebook or Contents of Datasets	_____	DARS # _____	Yes No
	_____	DARS # _____	Yes No
_____Journal Articles/Book Chapters	_____	DARS # _____	Yes No
	_____	DARS # _____	Yes No
_____Research/Technical Reports	_____	DARS # _____	Yes No
	_____	DARS # _____	Yes No
_____Final Reports	_____	DARS # _____	Yes No
	_____	DARS # _____	Yes No
_____Other	_____	DARS # _____	Yes No
	_____	DARS # _____	Yes No